Tree Species Detection on Satellite Imagery Data

Data Science Institute
COLUMBIA UNIVERSITY

Authors: Cyndi Chen, Kenny Wu, Yunshu Cai, Yunze Pan, Ziyan Liu
Industry Mentors: Harini Srinivasan, Gurkanwar Singh
Faculty Mentors: Sining Chen, Aayush Verma

Data Science Capstone
Project with

IBM

Background and Overview

Tree species information is important in several industrial applications, including tracking invasive species, understanding trees growth rate to plan trimming, and removing unhealthy trees by utility companies and railways, etc.

Our project aims to use remote-sensing satellite images and labeled tree species data with geo-locations to identify tree species. We used multiple ML algorithms, including deep learning techniques to build multiple tree species detection models.

Original Satellite Imagery — Tree Polygons and Species Info — Polygons Species Labeling

Pixel-based input

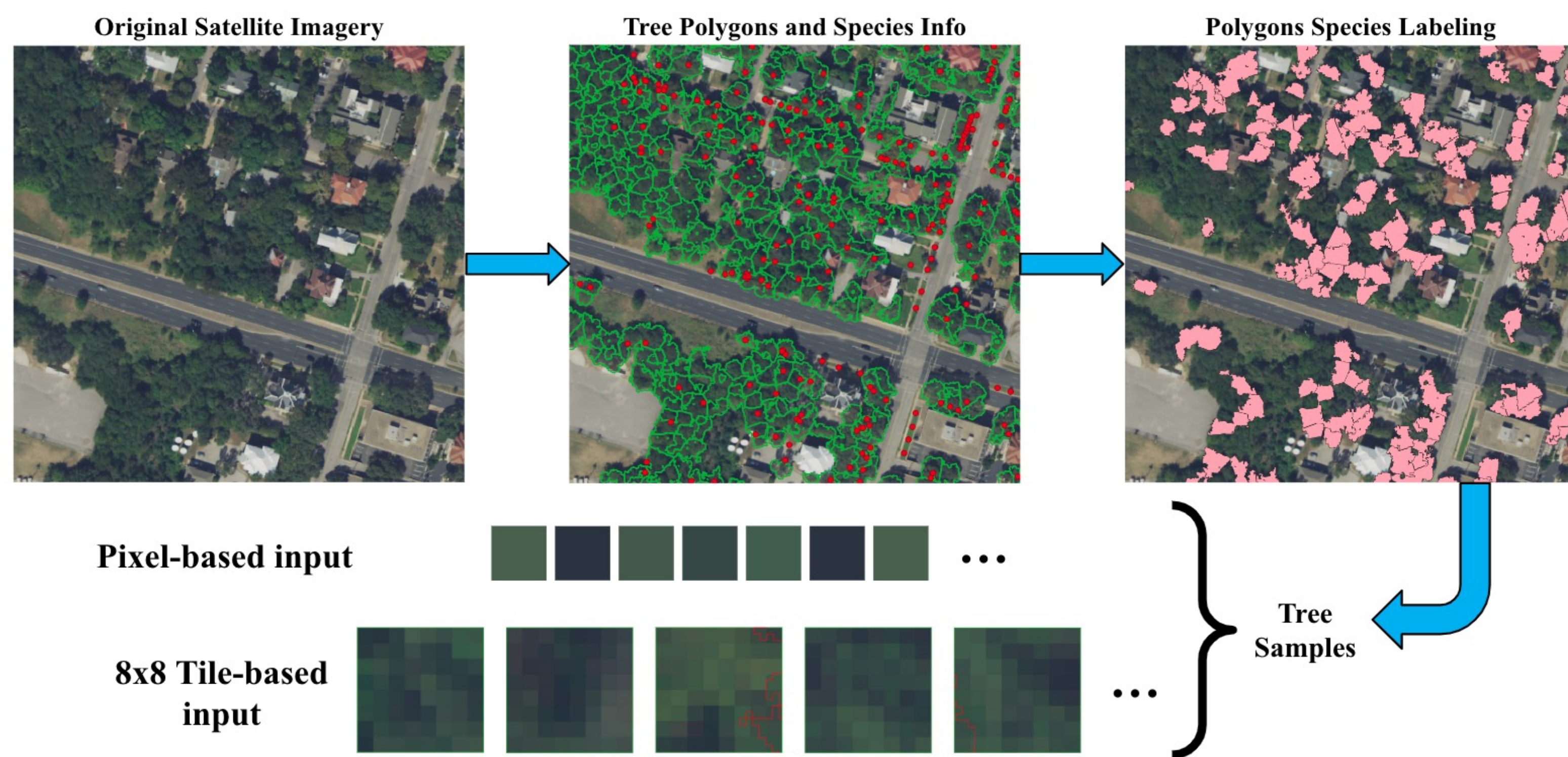8x8 Tile-based input

Tree Samples

Figure 1. Construction of the Tree Sample set from Austin Satellite Imagery

Data Analysis and Preparation

We visualized the satellite imagery in QGIS and processed it using Python.

- We firstly associated all the pixels with identified species.
- The core features for each pixel are the four bands (R,G,B,NIR) values, and we also calculated some aggregated features (Max, Standard deviation, Texture features).
- We finally produced input data that was pixel-based.
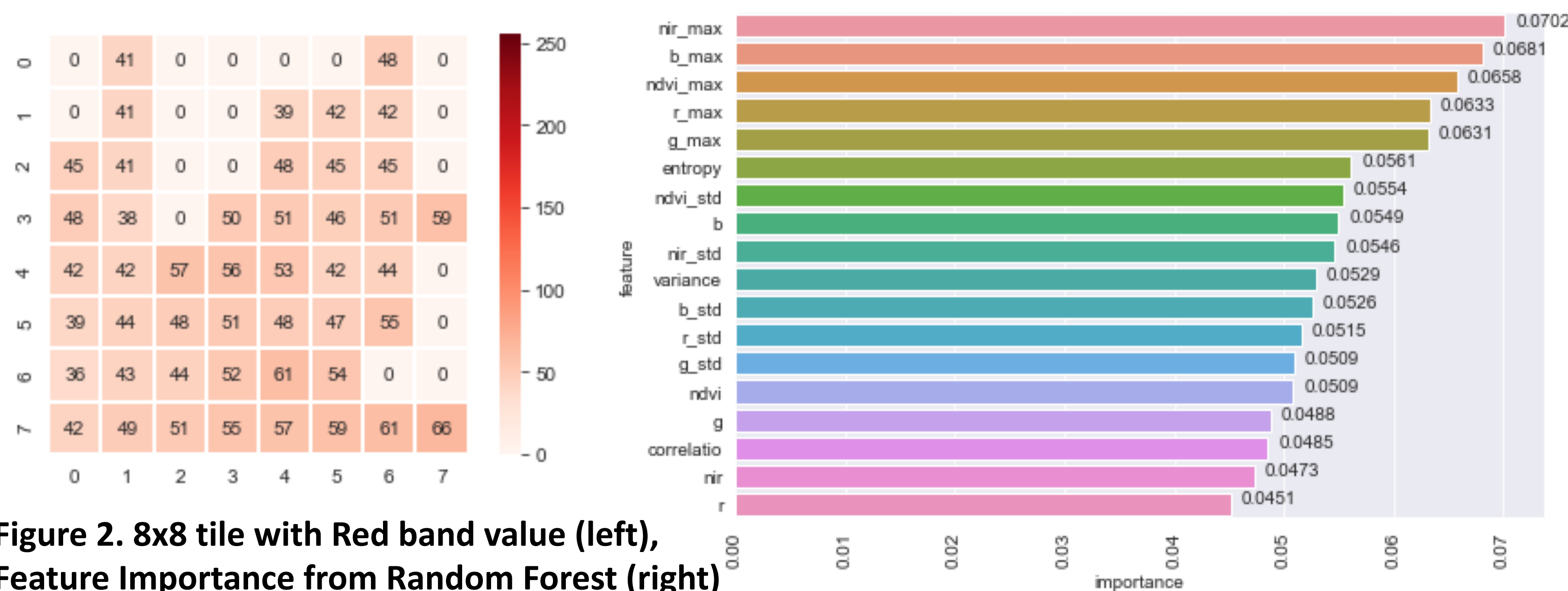- We also created tile-based input data using Python, where each tile is 8x8 pixels.

Figure 2. 8x8 tile with Red band value (left), Feature Importance from Random Forest (right)

Modeling

- **Pixel-based models:** We used Random Forest and SVM to classify our pixel-based data with 4 basic bands values and 14 aggregated features.
- **Tile-based models:** We focused on pre-trained CNN models like ResNet20, ResNet34 and DenseNet to classify the 8x8 pixel tiles. We firstly used the basic 4 bands features to find the best model and parameters. Then we added the Max aggregated features to refine the modelling process.
- **Model evaluation metrics:** Accuracy, Precision, Recall and F1 score for each species


```
========================Evaluation on Test Set========================

Accuracy on test set: 0.579
Cohen's kappa coefficient on test set: 0.462

====================Classification Report on Test Set====================

                      precision  recall  f1-score  sample num

               pecan       0.64    0.74      0.69        2035
            elm cedar       0.60    0.66      0.63        1661
     southern live oak       0.55    0.64      0.59        1623
   escarpment live oak       0.50    0.46      0.48        1105
          crape myrtle       0.50    0.17      0.25         467
            hackberry       0.34    0.13      0.19         372
          ashe juniper       0.53    0.24      0.34         229

            accuracy                          0.58        7492
           macro avg       0.52    0.43      0.45        7492
        weighted avg       0.56    0.58      0.56        7492
```
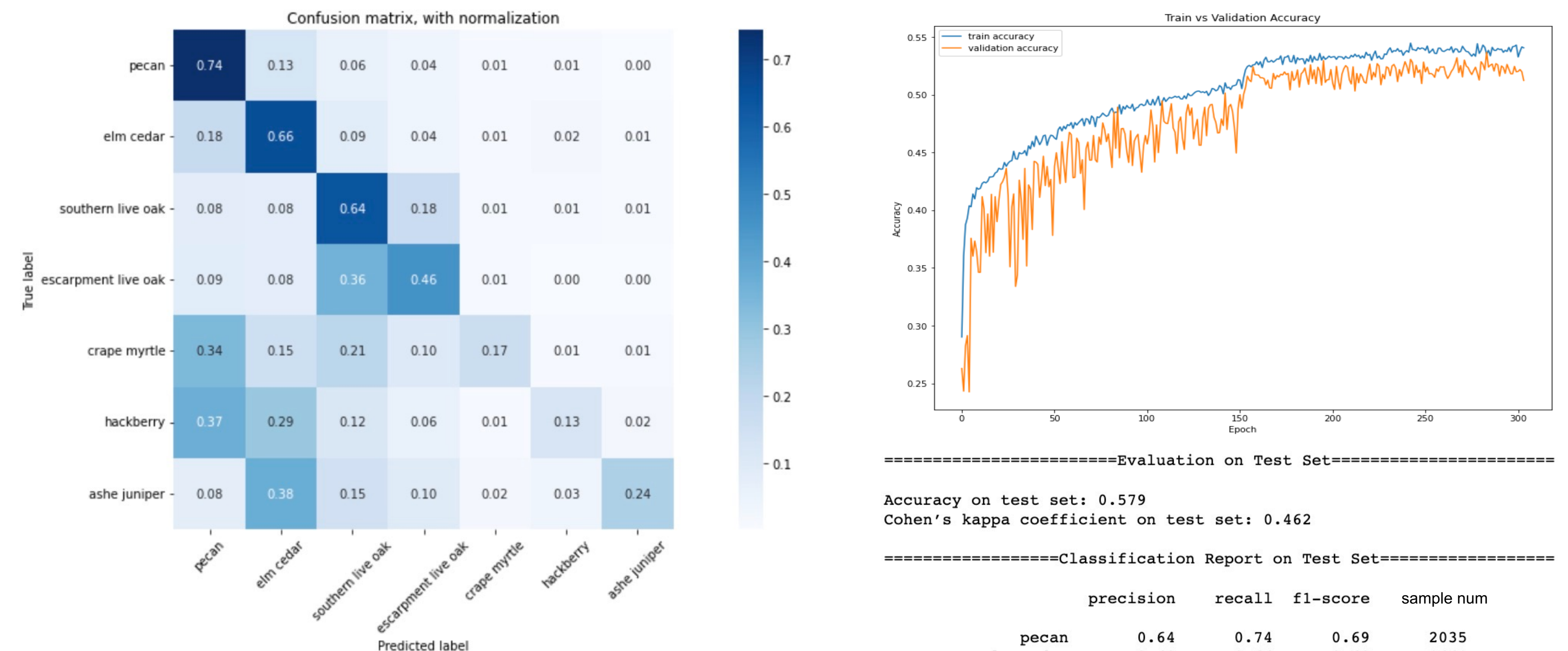
Figure 3. Confusion Matrix (left) and Evaluation Metrics (right) from ResNet20

Results and Conclusion

While some models show promising results, some others still need to be fine tuned.

- With total 18 features (R, G, B, NIR, NDVI, and aggregated features), Random Forest is the best performing model with 67% test accuracy.
- With the basic 4 bands features, ResNet20 is the best model with 58% test accuracy
- With more complex CNN-based model (U-Net), our classification accuracy could be improved.

Acknowledgments

We thank Harini Srinivasan and Gurkanwar Singh for guiding us throughout the project and providing feedback and suggestions on our paper.

References

Deur, M., Gašparović, M., & Balenović, I. - Tree species classification in mixed deciduous forests using very high spatial resolution satellite imagery and machine learning methods.

Natesan, S., Armenakis, C., & Vepakomma, U. (2020). Individual tree species identification using dense convolutional network (DenseNet) on multitemporal RGB images from UAV.